

1 **Emerging SARS-CoV-2 diversity revealed by rapid whole genome
2 sequence typing**

3

4 Ahmed M. Moustafa¹, PhD; Paul J. Planet^{1,2,3*}, PhD

5

6 **1. Division of Pediatric Infectious Diseases, Children's Hospital of Philadelphia,
7 Philadelphia, PA 19104, USA.**

8

9 **2. Department of Pediatrics, Perelman College of Medicine, University of
10 Pennsylvania, Philadelphia, PA 19104, USA.**

11

12 **3. Sackler Institute for Comparative Genomics, American Museum of Natural
13 History, New York, NY 10024, USA.**

14

15 **Emails**

16 **AMM: moustafaam@email.chop.edu**

17 **PJP: planetp@email.chop.edu**

18

19 **Phone**

20 **PJP: +1 215-590-1169**

21 ***Corresponding Author**

22

23 **Abstract**

24 **Background**

25 Discrete classification of SARS-CoV-2 viral genotypes can identify emerging strains and
26 detect geographic spread, viral diversity, and transmission events.

27 **Methods**

28 We developed a tool (GNUVID) that integrates whole genome multilocus sequence
29 typing and a supervised machine learning random forest-based classifier. We used
30 GNUVID to assign sequence type (ST) profiles to each of 69,686 SARS-CoV-2
31 complete, high-quality genomes available from GISAID as of October 20th 2020. STs
32 were then clustered into clonal complexes (CCs), and then used to train a machine
33 learning classifier. We used this tool to detect potential introduction and exportation
34 events, and to estimate effective viral diversity across locations and over time in 16 US
35 states.

36 **Results**

37 GNUVID is a scalable tool for viral genotype classification (available at
38 <https://github.com/ahmedmagds/GNUVID>) that can be used to quickly process tens of
39 thousands of genomes. Our genotyping ST/CC analysis uncovered dynamic local
40 changes in ST/CC prevalence and diversity with multiple replacement events in different
41 states. We detected an average of 20.6 putative introductions and 7.5 exportations for
42 each state. Effective viral diversity dropped in all states as shelter-in-place travel-
43 restrictions went into effect and increased as restrictions were lifted. Interestingly, our
44 analysis showed correlation between effective diversity and the date that state-wide
45 mask mandates were imposed.

46 **Conclusions**

47 Our classification tool uncovered multiple introduction and exportation events, as well as
48 waves of expansion and replacement of SARS-CoV-2 genotypes in different states.
49 Combined with future genomic sampling the GNUVID system could be used to track
50 circulating viral diversity and identify emerging clones and hotspots.

51

52 **Keywords**

53 *COVID-19; nomenclature; lineages; wgMLST; clonal complex; machine learning*

54 **Introduction**

55 Rapid sequencing of the SARS-CoV-2 pandemic virus has presented an
56 unprecedented opportunity to track the evolution of the virus and to understand the
57 emergence of a new pathogen in near-real time. During its explosive radiation and
58 global spread, the virus has accumulated enough genomic diversity that we are now
59 able to identify distinct lineages and track their spread in distinct geographic locations
60 and over time (Bedford, et al. 2020; Chen, et al. 2020; Deng, et al. 2020; Rambaut, et
61 al. 2020; Shen, et al. 2020; Worobey, et al. 2020). Phylogenetic analyses in
62 combination with rapidly growing databases (Shu and McCauley 2017; Rambaut, et al.
63 2020) have been instrumental in identifying distinct clades and tracing how they have
64 spread across the globe, as well as estimating calendar dates for the emergence of
65 certain clades (Bedford, et al. 2020; Deng, et al. 2020; Rambaut, et al. 2020; Worobey,
66 et al. 2020). This information is extremely useful in assessing the impact of early
67 measures to combat spread as well as identifying missed opportunities (Korber, et al.
68 2020; Worobey, et al. 2020).

69 Although reconstructing a robust phylogeny of viral variants is an intuitive
70 approach for viral classification, traditional phylogenetic approaches suffer from
71 problems with scalability. Building comprehensive phylogenetic trees for single
72 nucleotide polymorphism (SNP) based analysis of SARS-CoV-2 is already extremely
73 computationally expensive, and will become more and more difficult as hundreds of
74 thousands of sequences are added. Dividing the dataset into subsets of genomes
75 necessarily loses information and explanatory power. Because of this roadblock, our
76 goal was to develop a rapid way to categorize genomes that scales readily and leads to
77 as little information loss as possible. We saw an opportunity to combine our allele
78 identifying tool, WhatsGNU (Moustafa and Planet 2020b), with the Multilocus Sequence
79 Typing (MLST) approach (Maiden, et al. 1998) that has been widely used in bacterial
80 classification, tracking the emergence of new lineages, and associating specific
81 Sequence Types/Clonal Complexes (STs/CCs) with certain diseases. Our whole
82 genome MLST (wgMLST) approach rapidly assigns an allele number to each gene
83 nucleotide sequence in the virus's genome creating a sequence type (ST), which is

84 codified as the sequence of allele numbers for each of the ten genes in the viral
85 genome.

86 Here we show that this approach allows us to link STs into clearly defined clonal
87 complexes (CC) that are consistent with phylogeny and other SARS-CoV-2 typing
88 systems (Shu and McCauley 2017; Rambaut, et al. 2020). We show that assessment of
89 STs and CCs agrees with multiple introductions of the virus in certain US states. In
90 addition, we use temporal assessment of ST/CC diversity to uncover waves of
91 expansion and decline, and the apparent replacement of certain CCs with emerging
92 lineages in specific US states.

93

94 **Results and Discussion**

95 We developed the GNU-based Virus IDentification (GNUVID) system as a tool
96 that automatically assigns a number to each unique allele of the ten open reading
97 frames (ORFs) of SARS-CoV-2 (Wu, et al. 2020) (Figure 1A). GNUVID compressed the
98 696,860 ORFs in 69,686 high quality GISAID genomes (Supplementary Table 1) to
99 37,921 unique alleles in five minutes on a standard desktop, achieving 18-fold
100 compression and losing no information. To create an ST for each isolate GNUVID
101 automatically assigned 35,010 unique ST numbers based on their allelic profile
102 (Supplementary Table 1). We then used a minimum spanning tree (MST) to group STs
103 into larger taxonomic units, clonal complexes (CCs), which we define here as clusters of
104 >20 STs that are single or double allele variants away from a “founder”. Using the
105 goeBURST algorithm (Feil, et al. 2004; Francisco, et al. 2009) to build the MST and
106 identify founders, we found 154 CCs (Figure 1A and Supplementary Table 1).

107 A random forest classifier was then trained on 53,565 CC-labelled genomes. The
108 overall prediction statistics of the model were accuracy: 0.955, F-score: 0.950,
109 precision: 0.947, and recall: 0.964 (Figure 1B).

110 For any new query genome, GNUVID attempts to classify it first by exact
111 matching of the allelic profile to one of the other STs. If there is no exact match, the CC
112 for the query genome is predicted using the trained model. This query process saves
113 time and also allows each ORF to be typed and tallied individually (Figure 1C and 1D).

114 To show that CCs are mostly consistent with whole genome phylogenetic trees,
115 we mapped the 10 most common CC designations onto a maximum likelihood tree.
116 Members of the same CC usually grouped together in clades (Supplementary Figure 1).
117 To further validate our wgMLST classification system we compared it to the proposed
118 “dynamic lineages nomenclature” for SARS-CoV-2 (Rambaut, et al. 2020) and GISAID
119 clades naming system (Shu and McCauley 2017). A high percentage of CCs, 95.5%
120 (147/154) and 87.7% (135/154) of the CCs, had 90% of their genomes assigned to the
121 same GISAID clade and pangolin lineage, respectively, showing strong agreement
122 between these classification schemes (Supplementary Table 1). One limitation of our
123 classification strategy, as with many schemes that operate in real time, is that
124 paraphyletic groups can occur as a new ST arises from an older ST (e.g. CC258 and
125 CC768 emerged from CC255 and CC258 making CC255 and CC258 paraphyletic,
126 respectively) (Supplementary Figure 1). While this means that not all ST/CC groups will
127 be monophyletic, this property of the nomenclature may be helpful in gauging
128 emergence and replacement of an ancestral form.

129 When the global region of origin for each genome sequence was mapped to
130 each CC there was a strong association of later emerging CCs with certain
131 geographical locations, possibly reflecting relative containment after international travel
132 restrictions (Figure 2). To obtain an up-to-date picture of virus diversity in the US, we
133 analyzed 107,414 high coverage genomes (isolation dates between December 2019 to
134 October 20th 2020) from the GISAID (Supplementary table 1). There were 26,528
135 genomes isolated in the US in this dataset that belong to 87 of 154 CCs. Strikingly, 35%
136 of the genomes belong to CC258 (GISAID clade GH) and 75% of the genomes are
137 represented by just 10 CCs (CC4, 255, 256, 258, 300, 498 768, 3530, 10221, 21210)).
138 Moreover, 72% (63/87) of the CCs (representing 82% of the genomes) had the spike
139 D614G mutation that has been associated with increased spread (Korber, et al. 2020).
140 Interestingly, none of the US genomes were associated with any of the 12 CCs (26377,
141 26754, 27693, 27950, 28012, 28825, 29259, 29310, 30362, 31179, 31744 and 31942)
142 that have the spike protein A222V mutation (GISAID clade GV) that has been recently
143 associated with increased spread in the Europe (Hodcroft, et al. 2020). Ten of the 12
144 CCs with the A222V mutation were isolated only from Europe while the two other CCs

145 (27693 and 27950) had 2 genomes from Hong Kong and 6 from New Zealand,
146 respectively. This shows a strong association of this clade with Europe.

147 The relative proportions of STs or CCs isolated and sequenced may be a highly
148 biased statistic that is contingent upon where the isolate comes from, the decision to
149 sequence its genome, and the local capacity to sequence a whole genome. Certain
150 states (Washington, Texas and California) clearly sequenced more genomes than the
151 other states. Focusing on specific states may help to partially ameliorate this bias, and
152 we chose to focus on 16 states (Washington (WA), Texas (TX), California (CA),
153 Wisconsin (WI), New York (NY), Michigan (MI), Minnesota (MN), Louisiana (LA), Utah
154 (UT), Virginia (VA), Florida (FL), Oregon (OR), Massachusetts (MA), New Mexico (NM),
155 Maryland (MD), and Connecticut (CT)) with at least 200 genomes in the studied time
156 period, representing 92.6% (24,565/26,528) of all viral genomes available from the US.
157 The most common 20 CCs in these states, representing 86.5% (21261/24565) of the
158 genomes, are shown in Figure 2.

159 Because we included collection dates for each genomic sequence, we can use
160 STs and CCs to better understand the emergence and replacement of certain lineages
161 and viral diversity in geographical regions over time. Figure 3A and Supplementary
162 Figure 2 show temporal plots of the most common 20 CCs in 16 states. In WA, the
163 earlier introduction CC256 (GISAID clade S) was replaced by CC258 (GISAID clade
164 GH), perhaps by introduction from the East Coast or Europe (Bedford, et al. 2020;
165 Deng, et al. 2020). CC258 was then replaced by CC300 (GISAID clade GR) and
166 subsequently by CC498 (GISAID clade G).

167 In the neighboring state CA, a different pattern was seen in the early pandemic
168 where the lineage found early on in WA, CC256, only represented 20% of sequenced
169 genomes at its most prevalent (1st-15th March) while CC4 (GISAID clade L) was the
170 dominant variant, and was then replaced by CC258. Interestingly, a locally emerged
171 variant CC10221 (GISAID clade G), probably from CC498, increased in abundance
172 over time and then was likely exported to OR and NM (Supplementary Figure 2). A
173 similar pattern was seen in WI where a local variant CC13301 increased in abundance
174 over time and then appeared to spread to other states (NY, MI, MA and MN). In TX,
175 multiple diverse CCs persisted in the population until mid-July.

176 In NY, a different pattern is seen with CC258 being persistently dominant.
177 However, a more granular view of STs, not CCs, in New York shows a shifting
178 epidemiology with ST258 declining and the rise of closely related single and double
179 locus variants of ST258 reflecting local diversification (Supplementary Figure 3).

180 In MI, CC258 was the predominant strain until the summer when it gave way to a
181 more diverse group of isolates. Similarly, in states like VA, CT, NM and LA mostly one
182 predominant CC is seen over time, while in other states like UT, FL, OR, MA, MD and
183 MN a diverse pattern of multiple CCs was noticed (Supplementary Figure 2).

184 The expansions and contractions in the temporal plots over time could be due to
185 locally generated diversity (mutation) and/or introductions from other states or overseas.
186 To better understand the source of ST diversity over time, we calculated indices
187 reflecting effective circulating diversity as well as proportions of new STs in each state,
188 and inferred domestic or global introductions and exportations based on previous
189 observations in other locations or subsequent observations in other geographical
190 locations (Figure 3B, Table 1 and Supplementary Figure 4). To infer introductions, we
191 required that exactly the same ST was seen at least 10 days prior in some other
192 geographical location. For exportations we required an ST to be seen first in the state in
193 question at least 10 days prior to being seen anywhere else.

194 The results of this analysis showed distinct patterns in different states with
195 evidence supporting introductions usually outweighing evidence supporting exportations
196 (Table 1). Interestingly, NY has the highest number of putative exportations (n=26),
197 which was almost equal to the number of putative importations (n=25) potentially
198 reflecting its role as a hub driving the initial pandemic. In most states there was a high
199 amount of diversity that had no evidence of being introduced, which may signal
200 problems with sampling, or may signal that local mutation is a strong force in generating
201 diversity.

202 To understand the diversity within and between states, we calculated Hill
203 numbers for all genomes from each state and over time in each state (Figure 4A, Table
204 1). Hill numbers are a diversity metric used widely in ecological studies that express
205 effective diversity in units of sequence types, and they are less prone to biases
206 introduced by incomplete or biased sampling (Alberdi and Gilbert 2019). Recognizing

207 that our sample was not drawn from a systematically or evenly sampled dataset, we
208 chose to use a Hill number metric ($q=2$) that emphasizes abundant taxa in estimating
209 the effective diversity. Several other metrics such as the Shannon Index and a
210 normalized richness index were highly dependent on the number of sampled genomes
211 from each state. Hill numbers based on STs varied widely by state with TX showing the
212 highest diversity and MI showing the lowest (Figure 3B and 4 and Table 1).
213 Interestingly, there is a correlation ($R^2 = 0.1625$) between effective diversity and when a
214 state-wide mask mandate was imposed (Figure 4B).

215 Higher effective diversity may signal increased introduction of variants or
216 increased local generation of new sequence types, which in turn may signal more open
217 flow of virus into certain states or large circulating populations of virus able to mutate
218 and diversify, respectively. To attempt to discriminate between these processes we
219 calculated the effective diversity over time in each state and compared this to the
220 proportion of novel variants that were determined to be introductions (Figure 3B and
221 Supplementary Figure 4). In most states, initially high numbers of introductions were
222 followed by a drop in the relative proportion of introductions as states began to impose
223 restrictions in March. In some states the proportion of introductions also appears to
224 increase over the summer as states eased regulations. Interestingly effective diversity
225 also appeared to be correlated with peaks in the number of cases (Supplementary
226 Figure 5) in several states, especially New York, but more data will be needed to be
227 assessed to understand the connection between effective diversity and numbers of
228 cases reported.

229 While our wgMLST approach is rapid and robust it has several limitations.
230 Because a change in any allele creates a new ST our method may accumulate and
231 count “unnecessary” STs that have been seen only once or may be due to a
232 sequencing error. This is partially ameliorated by the use of the CC definition that allows
233 some variability amongst the members of a group, and the use of only high-quality
234 sequences. A large number of STs also may allow more granular approaches to
235 tracking new lineages. Another limitation is the stability of the classification system,
236 some virus genomes may be reassigned to new CCs as clones expand

237 epidemiologically, but this may also reflect a dynamic strength as circulating viruses
238 emerge and replace older lineages.

239 Perhaps most important limitation of our classification system is that it is limited
240 by the quality and extent of the database. This is also reflected in the major limitation
241 associated with the epidemiological and diversity inferences reported here. Uneven or
242 biased sampling could lead to both inaccurate statements of the direction or origin of
243 import/export events, and the source and quantification of diversity. The use of diversity
244 statistics that emphasize more predominant variants and address sampling bias such as
245 Hill numbers may help ameliorate this problem, but it seems clear that well-designed
246 sampling strategies are needed to confidently understand ecological dynamics for
247 SARS-CoV-2.

248

249 **Conclusion**

250 The genomic epidemiology of the 69,686 SARS-CoV-2 isolates studied here
251 show that 154 CCs have circulated globally and that more than half of these have been
252 dynamically spreading through the US population with waves of changing diversity. Our
253 tool (GNUVID) allows for fast sequence typing and clustering of whole genome
254 sequences in a rapidly changing pandemic. As illustrated above, this can be used to
255 temporally track emerging clones, identify the likely origin of viruses, and understand
256 circulating diversity.

257

258 **Materials and Methods**

259 All SARS-CoV-2 genomes (n=110,953) that were complete and have high
260 coverage were downloaded from GISAID (Shu and McCauley 2017) on October 20th
261 2020. Our wgMLST scheme was composed of all ten ORFs in the SARS-CoV-2
262 genome (Wu, et al. 2020). Genomes had to be at least 29,000 bp in length and have
263 fewer than 1% "N"s. The ten ORFs were identified in the genomes using blastn
264 (Altschul, et al. 1990) and any genome that had any ambiguity or degenerate bases
265 (any base other than A,T,G and C) in the ten open reading frames (ORF) was excluded.
266 The remaining 69,686 genomes (Supplementary table 1) were fed to the GNUVID tool
267 in a time order queue (first-collected to last-collected), which assigned an ST profile to

268 each genome. The identified STs by GNUVID were fed into the PHYLOViZ tool
269 (Nascimento, et al. 2017) to identify CCs at the double locus variant (DLV) level using
270 the goeBURST MST (Feil, et al. 2004; Francisco, et al. 2009). CCs were mapped back
271 to the STs using a custom script. Pie charts were plotted using a custom script. The sci-
272 kit learn implementation of Random Forest was then used to train a model. The model
273 was trained using 53,565 SARS-CoV-2 sequences from GISAID representing the 154
274 CCs. Briefly, the 53,565 genomes were aligned to MN908947.3(Wu, et al. 2020) to
275 generate a multiple sequence alignment using MAFFT's FFT-NS-2 algorithm(Katoh, et
276 al. 2002) (options: --add --keeplength). The 5' and 3' untranslated regions were masked
277 in the alignment file using a custom script. Variant positions were then called using snp-
278 sites (Page, et al. 2016) (options: -o -v). The 15,136 variant positions (features) matrix
279 of the 53,565 CC-labelled genomes were then one-hot encoded, in which each SNP is
280 replaced with a binary vector, and were used to train a random forest classifier in Scikit-
281 learn (Pedregosa, et al. 2011). The prediction capability of the model was evaluated
282 according to four statistics (accuracy, precision, recall and F-score).

283 To show the relationship between our typing scheme and phylogeny, we used a
284 Global phylogeny of SARS-CoV-2 sequences from GISAID (last accessed 2020-11-13).
285 The tree uses 99,160 high quality genomes(Lanfear and Mansfield. 2020). The tree and
286 the 10 most common CCs were visualized in iTOL (Letunic and Bork 2019). We
287 assigned a pangolin lineage (Rambaut, et al. 2020) (<https://github.com/hCoV-2019/pangolin>) and GISAID clade to each genome of the 53,565 genomes using the
288 metadata details available on GISAID. We then compared the composition of each CC
289 and calculated the percentage of the predominant clade/lineage in each CC
290 (Supplementary table 1).

292 A total of 107,414 genomes (Supplementary table 1), that were training examples
293 or assigned CCs and have date of isolation, were then used to analyze the number of
294 introductions and exportations. Putative introductions were defined as an exact ST that
295 was isolated somewhere else at least 10 days before the first date of isolation in the
296 state in question. Exportations were defined as STs that were first isolated in the state
297 in question and then isolated subsequently somewhere else at least 10 days later.

298 To compare diversity between the states and in each state over time, we
299 calculated the Simpson index (Simpson 1949). To measure effective diversity in units of
300 STs, we then transformed Simpson index (2H) to a Hill number (2D), which is the
301 multiplicative inverse of the Simpson index (Alberdi and Gilbert 2019). The dates of
302 state-wide mask mandates were the dates when face covering was required in indoor
303 public spaces and in outdoor public spaces when social distancing is not possible
304 (Abbott 2020; Allen 2020; Angell 2020; Baker 2020; Cuomo 2020; Edwards 2020; Evers
305 2020; Hogan 2020; Inslee 2020; Kunkel 2020; Lamont 2020; Northam 2020; Saunders
306 2020; Walz 2020; Whitmer 2020). The state-wide mandate dates used for WA, CA, TX,
307 WI, NY, MI, LA, FL, MN, NM, OR, MA, MD, VA, UT and CT are 6/26/20, 6/18/20, 7/3/20,
308 8/1/20, 4/17/20, 7/10/20, 7/11/20, no mandate, 7/25/20, 5/16/20, 7/13/20, 5/6/20,
309 7/31/20, 12/14/20, 11/9/20, and 4/17/20, respectively. The Hill number is described as
310 the effective number of STs (or CCs) of equally abundant STs (or CCs) that are needed
311 to give the same diversity (Hill 1973; Jost 2006). The plots for number of confirmed
312 cases in the 16 states were obtained from publicly available data in the Johns Hopkins
313 University dashboard (Dong, et al. 2020).

314 The GNUVID database will be updated regularly with new added high-quality
315 genomes from GISAID (Shu and McCauley 2017). Commands used are in
316 Supplementary Methods. All the scripts are available from the authors and
317 <https://github.com/ahmedmagds/GNUVID> (Moustafa and Planet 2020a). GNUVID can
318 be installed through Bioconda (Grüning, et al. 2018).

319

320 **Availability of data and material**

321 The compressed database and the trained model from our quality controlled genomes
322 are available from the corresponding author and available online for download
323 (Moustafa and Planet 2020a). The compressed database will be updated regularly on
324 <https://github.com/ahmedmagds/GNUVID>. Source code for GNUVID can be found in its
325 most up-to-date version here, <https://github.com/ahmedmagds/GNUVID>, under the
326 GNU General Public License. All scripts are available from the authors.

327

328 **Conflict of interest**

329 The authors declare that they have no competing interests

330

331 **Authors' contributions**

332 Conceptualization: AMM, PJP; Coding: AMM; Writing – Reviewing and Editing: AMM,
333 PJP.

334

335 **Acknowledgements**

336 We would like to thank Ms Lidiya Denu, Dr Michael Silverman at the Children's Hospital
337 of Philadelphia, Mr Apurva Narechania at the American Museum of Natural History, and
338 Dr. Joshua Mell Chang at Drexel University for helpful comments and discussion. We
339 would like to thank the Global Initiative on Sharing All Influenza Data (GISAID) and
340 thousands of contributing laboratories for making the genomes publicly available. A full
341 acknowledgements table is available at <https://github.com/ahmedmagds/GNUVID>. This
342 work was supported by the National Institute of Allergy and Infectious Diseases at the
343 National Institutes of Health (1R01AI137526-01 and 1R21AI144561-01A1 to A.M.M.
344 and P.J.P. and R01NR015639 to P.J.P.) and the Cystic Fibrosis Foundation
345 (PLANET19G0 to A.M.M. and P.J.P.).

346

347 **Table 1.** Number of Genomes, Sequence Types, Simpson index, Hill Number,
348 introductions and exportations for 16 US states.

349

State	Genomes (STs)	Simpson Index (² H)	Hill Number (² D)	Non- introductions	Introductions (US)	Exportations
WA	3960 (1887)	0.987	77	1817	44 (26)	19
TX	2167 (1299)	0.997	319	1258	31 (16)	17
CA	1984 (1236)	0.997	296	1173	35 (19)	7
NY	1483 (825)	0.960	25	766	25 (9)	26
MN	1107 (522)	0.988	81	470	29 (17)	12
WI	954 (574)	0.993	147	529	26 (15)	8
VA	908 (543)	0.994	165	511	18 (13)	4
LA	850 (416)	0.988	85	397	10 (10)	1
MI	795 (416)	0.889	9	384	16 (5)	9
FL	750 (519)	0.995	215	474	29 (18)	6
OR	531 (343)	0.995	190	320	19 (14)	5
UT	350 (216)	0.992	123	204	8 (4)	2
MA	336 (170)	0.940	17	144	17 (12)	2
MD	196 (145)	0.987	76	134	8 (4)	2
NM	162 (109)	0.987	80	103	3 (1)	0
CT	154 (101)	0.964	28	84	12 (8)	0

350 **Figure Legends**

351 **Figure 1.** Workflow for the GNUVID tool and its compression technique. **A.**

352 **Compression and classification.** The tool starts by compressing the database of the
353 10 ORFs of each of the SARS-CoV-2 genomes to only include a unique sequence for
354 each allele type. The tool then uses a whole genome multilocus sequence typing
355 (wgMLST) approach by assigning an allele number to each gene nucleotide sequence
356 in the virus's genome creating a sequence type (ST) which is codified as the sequence
357 of allele numbers for each of the ten genes in the viral genome. The STs are then linked
358 into clearly defined clonal complexes (CCs) using goeBURST . **B. Training a machine**
359 **learning classifier.** The CC-labelled genomes are then aligned to the SARS-CoV-2
360 reference genome (MN908947.3) and single nucleotide polymorphisms (SNPs) are
361 called. The SNP matrix is then one-hot encoded and used to train a random forest
362 classifier. The training followed a 5-fold cross-validation approach to assess the
363 prediction capabilities of GNUVID according to four statistics (accuracy, precision, recall
364 and F-score). TP, TN, FP and FN are true positives, true negatives, false positives and
365 false negatives, respectively. **C. New Genome classification by exact matching or**
366 **prediction.** GNUVID first tries to match each of the 10 ORFs from a query SARS-CoV-
367 2 genome to an exact match in the compressed database to define an ST, and matches
368 that to any associated CC. If no exact match is found due to novelty or ambiguity in any
369 of the 10 ORFs, the query genome is aligned to the reference, one-hot encoded and a
370 CC is predicted by the trained classifier. A report is then created showing the allele
371 number of each ORF, ST, CC and a probability of membership in the CC. **D. Map of**
372 **SARS-CoV-2 virus genome showing the length in base pairs (bp) of the ten ORFs and**
373 **numbers of alleles in the current database 69,686 isolates.** The majority of the identified
374 37,921 unique alleles (69%) are for ORF1ab which represents 71% of the genome
375 length. Strikingly, the two highest ratios (number of alleles/ORF length) are for the
376 nucleocapsid protein (2.2) and ORF3a (2.1) while the spike protein had a ratio of 1.32.
377
378 **Figure 2. Global SARS-CoV-2 Diversity.** Minimum spanning tree from goeBURST of
379 the 35,010 Sequence Types (STs) showing the 154 Clonal Complexes (CCs) identified
380 in the dataset. Only the most common 20 CCs in the 16 states are shown in black. The

381 pie charts show the percentage of genomes from the different geographic regions in
382 each CC.

383

384 **Figure 3. SARS-CoV-2 diversity in 6 states over time. A.** Temporal Plots of
385 circulating Clonal Complexes and corresponding GISAID clade in parentheses in six
386 different states (Washington (WA), California (CA), Wisconsin (WI), Texas (TX), New
387 York (NY) and Michigan (MI)). The visualizations were limited to the 20 most common
388 CCs. **B.** Diversity of Sequence Types (STs) in the six states over time are represented
389 for each 2-week time period in the following ratios: 1. Effective diversity (Hill number
390 equivalent (2D) of Simpson index (2H)) (red) 2. Number of STs new to a state that were
391 previously isolated and sequenced outside a state divided by the number of STs not
392 seen previously in a state (blue).

393

394 **Figure 4. Effective Diversity of Sequence Types (STs) in 16 states. A.** The Hill
395 number equivalent (2D) of Simpson index (2H), is on the y-axis. Total number of
396 genomes sequenced on the x-axis. **B.** Effective diversity (Hill number 2D) plotted
397 against the week when state-wide mask mandate was imposed. Florida (FL) has no
398 mask mandate so it was plotted at the end of the y-axis. The 16 different states are
399 Washington (WA), California (CA), Wisconsin (WI), Texas (TX), New York (NY),
400 Michigan (MI), Utah (UT), Virginia (VA), Florida (FL), Oregon (OR), Massachusetts
401 (MA), New Mexico (NM), Maryland (MD), Connecticut (CT), Minnesota (MN) and
402 Louisiana (LA).

403

404 **Additional files**

405 **Additional file 1:** Supplementary Methods and Figures.

406 **Additional file 2:** Table S1. GNUVID Database Strains Report Table.

407

408 **References**

409 Executive Order No. GA-29 relating to the use of face coverings during the COVID- 19 disaster in
410 Texas. [Internet]. 2020. Available from:

411 [https://open.texas.gov/uploads/files/organization/opentexas/EO-GA-29-use-of-face-coverings-
412 during-COVID-19-IMAGE-07-02-2020.pdf](https://open.texas.gov/uploads/files/organization/opentexas/EO-GA-29-use-of-face-coverings-during-COVID-19-IMAGE-07-02-2020.pdf)

413 Alberdi A, Gilbert MTP. 2019. A guide to the application of Hill numbers to DNA-based diversity
414 analyses. *Mol Ecol Resour* 19:804-817.

415 OHA Announces New Mask Requirements Website. Oregon Health Authority. [Internet]. 2020.
416 Available from: <https://www.oregon.gov/oha/ERD/Pages/OHA-Announces-New-Mask->
417 <Requirements-Website.aspx>

418 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
419 *Journal of Molecular Biology* 215:403-410.

420 Guidance for the use of face coverings. California Department of Public Health. [Internet]. 2020.
421 Available from:
422 https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/COVID-19/Guidance-for-Face-Coverings_06-18-2020.pdf

423 Order requiring face coverings in public places where social distancing is not possible.
424 Commonwealth of Massachusetts. [Internet]. 2020. Available from:
425 <https://www.mass.gov/doc/may-1-2020-masks-and-face-coverings/download>

426 Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang ML, Nalla A, Pepper
427 G, Reinhardt A, Xie H, et al. 2020. Cryptic transmission of SARS-CoV-2 in Washington state.
428 Science.

429 Chen Z-w, Li Z, Li H, Ren H, Hu P. 2020. Global genetic diversity patterns and transmissions of
430 SARS-CoV-2. *medRxiv*:2020.2005.2005.20091413.

431 Executive Order No. 202.17. State of New York. [Internet]. 2020. Available from:
432 https://www.governor.ny.gov/sites/governor.ny.gov/files/atoms/files/EO_202.17.pdf

433 Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria NR, Wang C, Yu G, Bushnell B, Pan CY,
434 et al. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern
435 California. *Science* 369:582-587.

436 Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real
437 time. *The Lancet Infectious Diseases* 20:533-534.

438 Proclamation number 89 JBE 2020. State of Louisiana [Internet]. 2020. Available from:
439 <https://gov.louisiana.gov/assets/Proclamations/2020/89-JBE-2020.pdf>

440 Emergency Order 1. State of Wisconsin. [Internet]. 2020. Available from:
441 <https://evers.wi.gov/Documents/COVID19/EmO01-FaceCoverings.pdf>

442 Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. 2004. eBURST: Inferring Patterns of
443 Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence
444 Typing Data. *Journal of Bacteriology* 186:1518.

445 Francisco AP, Bugalho M, Ramirez M, Carriço JA. 2009. Global optimal eBURST analysis of
446 multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* 10:152.

447 Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, The
448 Bioconda T. 2018. Bioconda: sustainable and comprehensive software distribution for the life
449 sciences. *Nature Methods* 15:475-476.

450 Hill MO. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*
451 54:427-432.

452 Hodcroft EB, Zuber M, Nadeau S, Crawford KHD, Bloom JD, Veesler D, Vaughan TG, Comas I,
453 Candelas FG, Stadler T, et al. 2020. Emergence and spread of a SARS-CoV-2 variant through
454 Europe in the summer of 2020. *medRxiv*:2020.2010.2025.20219063.

456 Order of the Governor of the State of Maryland Number 20-07-29-01. [Internet]. 2020.
457 Available from: <https://governor.maryland.gov/wp-content/uploads/2020/07/Gatherings-10th-AMENDED-7.29.20.pdf>

459 Proclamation by the Governor of Washington Amending and Extending Proclamation 20-05
460 [Internet]. 2020. Available from: https://www.governor.wa.gov/sites/default/files/proc_20-60.pdf

461 Jost L. 2006. Entropy and diversity. *Oikos* 113:363-375.

462 Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
463 sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059-3066.

464 Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfaltrerer W, Hengartner N, Giorgi EE,
465 Bhattacharya T, Foley B, et al. 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that
466 D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182:812-827 e819.

467 Public Health Order. New Mexico Department of Health. [Internet]. 2020. Available from:
468 https://www.governor.state.nm.us/wp-content/uploads/2020/05/05_15_2020_PHO.pdf

469 Executive order NO. 7BB. State of Connecticut. [Internet]. 2020. Available from:
470 <https://portal.ct.gov/-/media/Office-of-the-Governor/Executive-Orders/Lamont-Executive-Orders/Executive-Order-No-7BB.pdf?la=en>

471 Lanfear R, Mansfield. R. 2020. roblanf/sarscov2phylo: 13-11-20. doi:10.5281/zenodo.4289383.
472 In.

473 Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new
474 developments. *Nucleic Acids Res* 47:W256-W259.

475 Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant
476 DA, et al. 1998. Multilocus sequence typing: A portable approach to the identification of clones
477 within populations of pathogenic microorganisms. *PNAS* 95:3140-3145.

478 Moustafa AM, Planet PJ. ahmedmagds/GNUVID: GNUVID v2.0: Globally circulating clonal
479 complexes as of 2020-10-20 [Internet]. 2020a. Available from: DOI:10.5281/zenodo.4313855.
480 Published online Dec 9.

481 Moustafa AM, Planet PJ. 2020b. WhatsGNU: a tool for identifying proteomic novelty. *Genome
482 Biology* 21:58.

483 Nascimento M, Sousa A, Ramirez M, Francisco AP, Carrico JA, Vaz C. 2017. PHYLOViZ 2.0:
484 providing scalable data integration and visualization for multiple phylogenetic inference
485 methods. *Bioinformatics* 33:128-129.

486 Executive Order Number 72. Commonwealth of Virginia [Internet]. 2020. Available from:
487 [https://www.governor.virginia.gov/media/governorvirginiagov/executive-actions/EO-72-and-Order-of-Public-Health-Emergency-Nine-Common-Sense-Surge-Restrictions-Certain-Temporary-Restrictions-Due-to-Novel-Coronavirus-\(COVID-19\).pdf](https://www.governor.virginia.gov/media/governorvirginiagov/executive-actions/EO-72-and-Order-of-Public-Health-Emergency-Nine-Common-Sense-Surge-Restrictions-Certain-Temporary-Restrictions-Due-to-Novel-Coronavirus-(COVID-19).pdf)

488 Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid
489 efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2:e000056.

490 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P,
491 Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine
492 Learning Research* 12 2825-2830.

493 Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A
494 dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat
495 Microbiol.*

500 Utah State Public Health Order 2020-27. [Internet]. 2020. Available from: https://coronavirus-download.utah.gov/Health/UPOH_2020-27_Updated%20Statewide_COVID-19_Restrictions.pdf

501

502 Shen L, Dien Bard J, Biegel JA, Judkins AR, Gai X. 2020. Comprehensive Genome Analysis of

503 6,000 USA SARS-CoV-2 Isolates Reveals Haplotype Signatures and Localized Transmission

504 Patterns by State and by Country. *Frontiers in Microbiology* 11.

505 Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to

506 reality. *Euro Surveill* 22.

507 Simpson EH. 1949. Measurement of Diversity. *Nature* 163:688-688.

508 Emergency Executive Order 20-81. Requiring Minnesotans to Wear a Face Covering in Certain

509 Settings to Prevent the Spread of COVID-19. [Internet]. 2020. Available from:

510 <https://www.leg.mn.gov/archive/execorders/20-81.pdf>

511 Executive Order No. 2020-147. State of Michigan. [Internet]. 2020. Available from:

512 https://www.michigan.gov/documents/whitmer/EO_2020-147_696551_7.pdf

513 Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, Rambaut A, Suchard MA, Wertheim

514 JO, Lemey P. 2020. The emergence of SARS-CoV-2 in Europe and North America. *Science*.

515 Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, et al. 2020. A

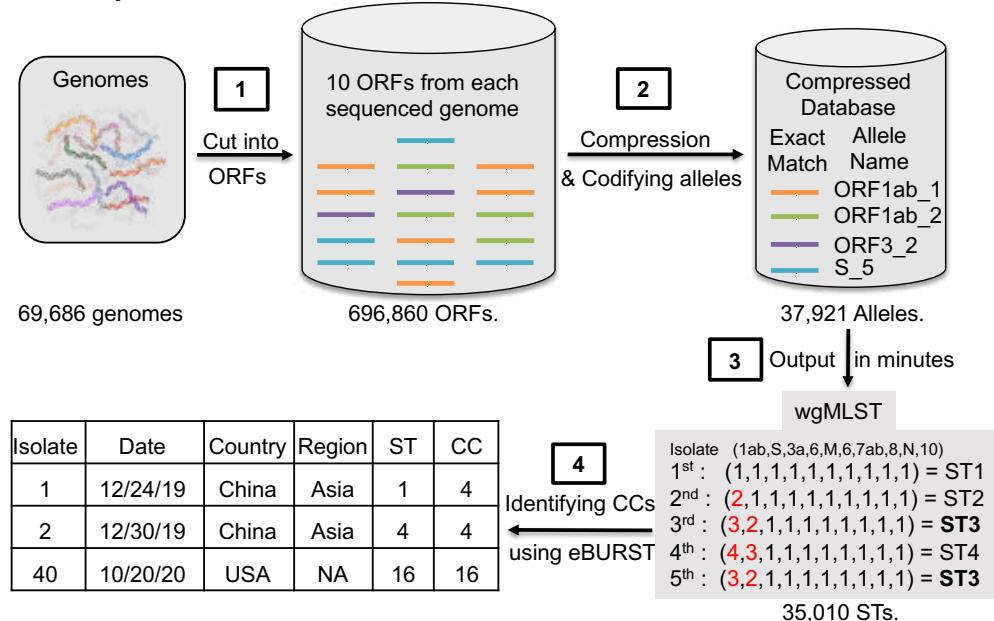
516 new coronavirus associated with human respiratory disease in China. *Nature* 579:265-269.

517

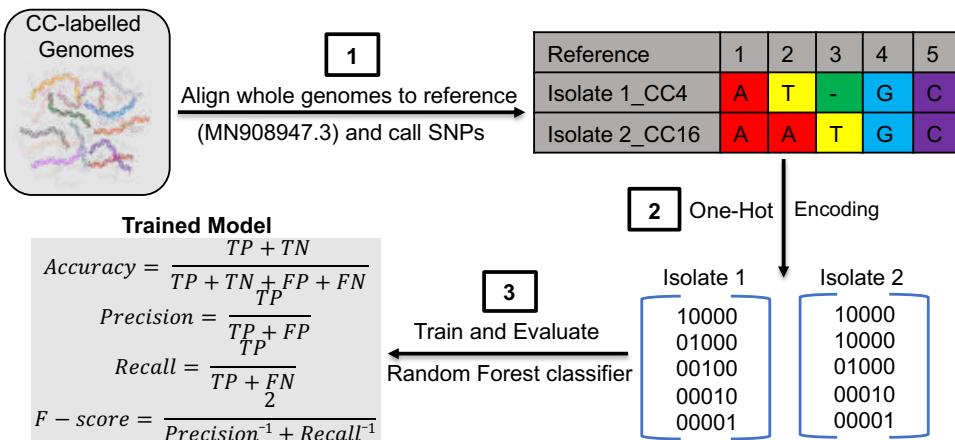
518

519 **Figure 1**

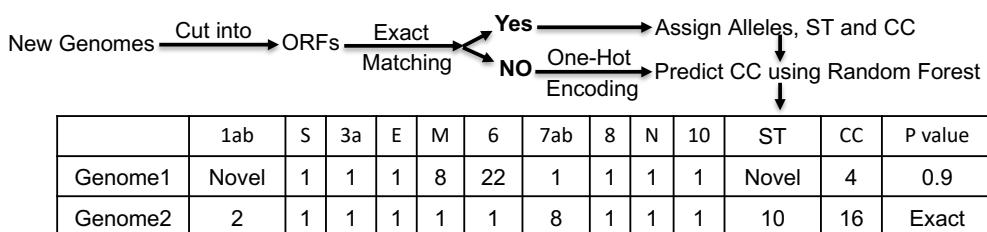
A. Compression and Classification



B. Training a Machine Learning Classifier



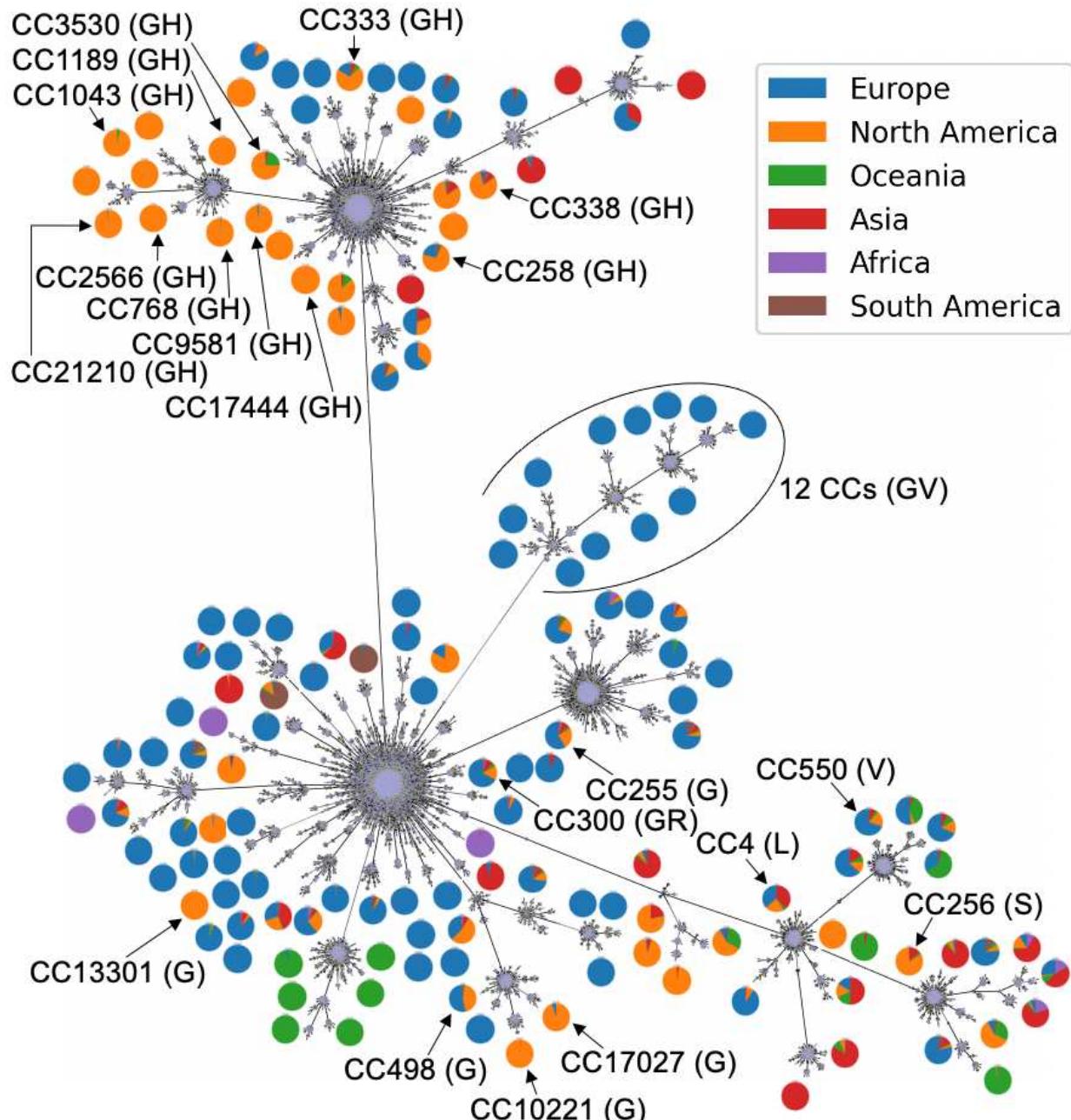
C. New genome classification by exact matching or prediction



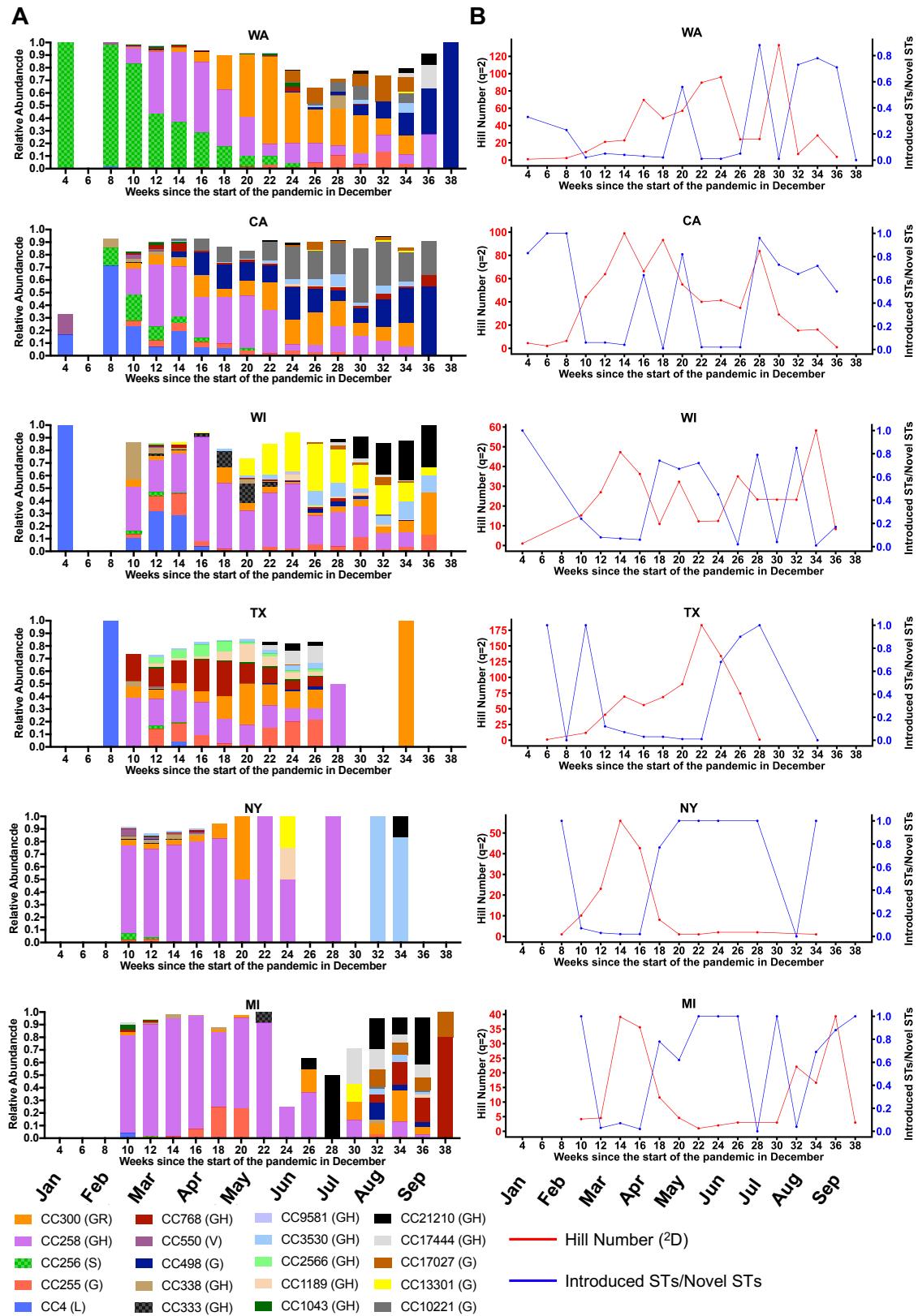
D. Numbers of alleles for each ORF

	1ab	S	3a	E	M	6	7a	8	N	10		
Length (bp)	21290		3822	828	228	669	186	366	366	1260	117	
Number of Alleles	26190		5063	1710	165	602	191	496	602	2760	142	

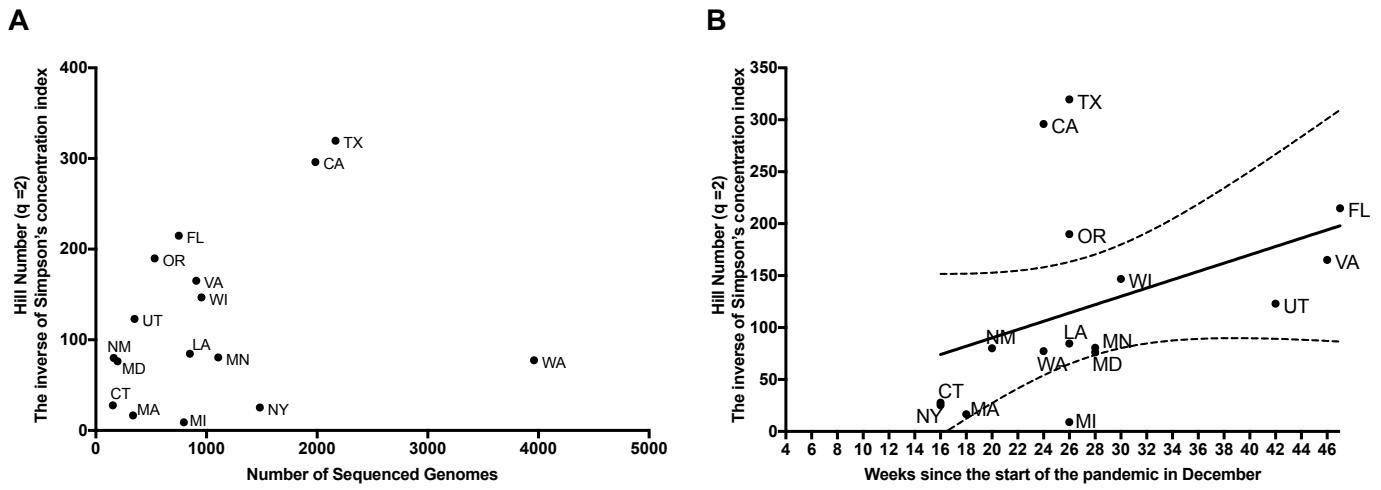
521 **Figure 2**



524 **Figure 3**



526 **Figure 4**



527
528